Hacia un método computacional para detectar léxico contrastivo

Damián Eliel Aleman

13 de noviembre de 2017

Temario

1. Introducción
2. Trabajo Previo
3. Datos: Extracción y procesamiento
   - Búsquedas geolocalizadas
   - Tokenización y normalización
4. Métricas para detectar léxico contrastivo
   - Primeras métricas: MaxDif y $MaxDif_g$
5. Análisis de las palabras contrastivas encontradas
   - Caracterización de las palabras identificadas como contrastivas
   - Validación estadística
6. Conclusiones y trabajo futuro

Qué es una palabra contrastiva

Se dice que una palabra es *contrastiva* cuando la frecuencia de uso en distintas regiones es muy diferente.

Ejemplos palabras contrastivas Argentina - España

- "che"
- "metegol"

Ejemplos palabras contrastivas dentro de Argentina

- "gurisada"

¿Para qué sirve conocer las palabras contrastivas?

Motivación de un método computacional para detectar léxico contrastivo

¿Cómo se conocían las palabras contrastivas? Mediante encuestas.

- Es costoso de realizar
- Muy difícil de hacer de forma balanceada en distintas regiones de un país o de un continente.
- Más difícil es encuestar a una gran cantidad de personas.
- Se basan en el conocimiento *a priori*

Temario

1. Introducción
2. Trabajo Previo
3. Datos: Extracción y procesamiento
   - Búsquedas geolocalizadas
   - Tokenización y normalización
4. Métricas para detectar léxico contrastivo
   - Primeras métricas: MaxDif y $MaxDif_g$
5. Análisis de las palabras contrastivas encontradas
   - Caracterización de las palabras identificadas como contrastivas
   - Validación estadística
6. Conclusiones y trabajo futuro

Web como un corpus

Kilgariff comentaba la riqueza de la Web para acceder a una información, antes impensada. Principalmente se hacían estudios estimando la frecuencia de una palabra viendo la cantidad de resultados de las consultas en un motor de búsqueda.

Ventajas:

- Gratuito
- Gran cantidad de datos
- Disponible e inmediato

Utilidades:

- Corrección de ortografía
- Traducción de frases
- Estimar el tamaño de la Web

Cita

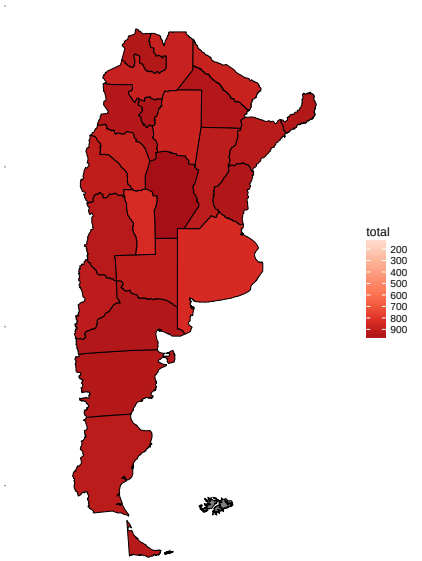
"La web es un corpus sucio, pero el uso esperado es mucho más frecuente que lo que puede considerarse como ruido." [KG03, p. 342]

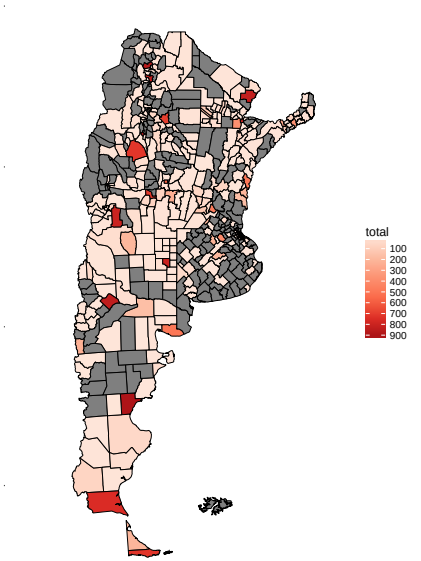
Temario

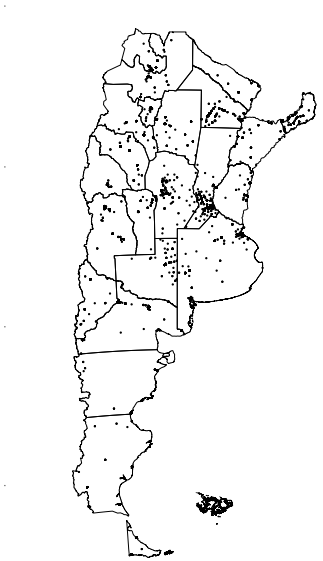
1. Introducción
2. Trabajo Previo
3. Datos: Extracción y procesamiento
   - Búsquedas geolocalizadas
   - Tokenización y normalización
4. Métricas para detectar léxico contrastivo
   - Primeras métricas: MaxDif y $MaxDif_g$
5. Análisis de las palabras contrastivas encontradas
   - Caracterización de las palabras identificadas como contrastivas
   - Validación estadística
6. Conclusiones y trabajo futuro

Búsquedas geolocalizadas

La búsqueda geolocalizada es una herramienta que nos da la posibilidad de obtener tuits generados en un área geográfica particular. Para esto, primero intenta buscar tuits cuyas coordenadas sean las buscadas.







Se realizaron búsquedas por cada provincia con centro en las coordenadas de los departamentos de la misma y con un radio de 20 millas.
Nos quedamos con los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia.

Tokenización y normalización del texto

Tokenización

Se consideró una palabra a las secuencias de caracteres formados únicamente por letras. Por lo tanto se eliminaron las menciones con @, los hashtags y links entre otros. Decidimos ignorar estos términos ya que no tienen interés lingüístico y agregarían mucho ruido a los datos.

Normalización

Todas las letras se convirtieron a letra minúscula y las palabras con más de tres letras iguales de forma consecutiva se redujeron para que solo tengan tres repeticiones. De esta forma, el término *padreeeee* y *padreeee* fueron reducidos a una única unidad léxica (*padreee*). Esto se hizo con la librería *TweetTokenizer de NLTK*.

Temario

1. Introducción
2. Trabajo Previo
3. Datos: Extracción y procesamiento
   - Búsquedas geolocalizadas
   - Tokenización y normalización
4. Métricas para detectar léxico contrastivo
   - Primeras métricas: MaxDif y $MaxDif_g$
5. Análisis de las palabras contrastivas encontradas
   - Caracterización de las palabras identificadas como contrastivas
   - Validación estadística
6. Conclusiones y trabajo futuro

Primeras métricas: MaxDif

Para cada palabra ω y cada par de provincias p_1 y p_2, podemos calcular el cociente entre la frecuencia máxima de ω en ambas provincias y la frecuencia mínima:

$$maxDif(\omega, p_1, p_2) = \frac{f_{max}(\omega, p_1, p_2)}{f_{min}(\omega, p_1, p_2)} \quad (1)$$

Desventajas:

1. Un valor para cada par de provincias.
2. No se considera la dispersión de los valores en todas las provincias.

Primeras métricas: $MaxDif_g$

Considerando las frecuencias de una palabra ω sobre todas las provincias, definimos:

$$maxDif_g(\omega) = \frac{f'_{max}(\omega)}{f'_{min}(\omega)} \tag{2}$$

donde $f'_{max}(\omega)$ es la frecuencia máxima de la palabra ω entre las frecuencias de todas las provincias y $f'_{min}(\omega)$ es la frecuencia mínima distinta de 0.
Con $maxDif_g$ se soluciona 1, pero no 2.

La entropía de la información

La entropía nos brinda un valor que indica qué tan uniforme es la distribución de las frecuencias de cada palabra.

Valor de la información

Zanette y Montemurro definieron al *valor de la información de una palabra* como

$$\Delta I_w(\omega) = p(\omega)\,(\widehat{H}(\omega) - H(\omega)) = p(w)\,\Delta H(\omega) \tag{3}$$

siendo $p(\omega)$ la frecuencia total de la palabra en el texto.

Valor contrastivo sobre las palabras

Valor contrastivo sobre las palabras

$$I_w(\omega) = norm_w(\omega) \cdot (\widehat{H}_w(\omega) - H_w(\omega)) \tag{4}$$

donde $norm_w$ sirve para normalizar sobre la cantidad de ocurrencias de la palabra.

Agregando la cantidad de personas que mencionan cada palabra

Hay palabras que pueden tener una frecuencia alta debido a pocas personas que las mencionan constantemente.

Valor contrastivo sobre las personas

$$I_p(\omega) = norm_p(\omega) \cdot (\widehat{H}_p(\omega) - H_p(\omega)) \tag{5}$$

donde $norm_p$ sirve para normalizar sobre la cantidad de personas que mencionan la palabra.

Temario

1. Introducción
2. Trabajo Previo
3. Datos: Extracción y procesamiento
   - Búsquedas geolocalizadas
   - Tokenización y normalización
4. Métricas para detectar léxico contrastivo
   - Primeras métricas: MaxDif y $MaxDif_g$
5. Análisis de las palabras contrastivas encontradas
   - Caracterización de las palabras identificadas como contrastivas
   - Validación estadística
6. Conclusiones y trabajo futuro

Temario

1. Introducción
2. Trabajo Previo
3. Datos: Extracción y procesamiento
   - Búsquedas geolocalizadas
   - Tokenización y normalización
4. Métricas para detectar léxico contrastivo
   - Primeras métricas: MaxDif y $MaxDif_g$
5. Análisis de las palabras contrastivas encontradas
   - Caracterización de las palabras identificadas como contrastivas
   - Validación estadística
6. Conclusiones y trabajo futuro


Adam Kilgarriff and Gregory Grefenstette.
Introduction to the special issue on the web as corpus.
Computational linguistics, 29(3):333–347, 2003.